UNIVERSITY OF TWENTE.

Health-related Quality of Life relating to Chronic Obstructive Pulmonary Disease

A psychometric Analysis of a new Disease-Specific Instrument

Nadine Herzog
s1155962
Faculty: Behavioral Science
Degree Program: Psychology
First Supervisor: Muirne C. S. Paap
Second Supervisor: Stéphanie van den Berg

15.06.2015

Table of Contents

List of Abbreviations

BPQ	Breathing Problems Questionnaire
CAT	Computer Adaptive Test
COPD	Chronic Obstructive Pulmonary Disease
COPD-SIB	Chronis Obstructive Pulmonary Disease specific Item Bank
CRQ	Chronic Respiratory Questionnaire
HRQoL	Health-related Quality of Life
IRT	Item Response Theory
MRF-26	26-item Maugeri Respiratory Failure Questionnaire
MST	Medisch Spectrum Twente
NHP	Nottingham Health Profile
QoL	Quality of Life
QoLRIQ	Quality of Life for Respiratory Illness Questionnaire
PROMIS	Patient Reported Outcomes Measurment Information System
SF-36	Short Form 36-item Questionnaire
SGRQ	St. George Respiratory Questionnaire
SGRQ-C	St. George Respiratory Questionnaire for COPD patients
SIP	Sickness Impact Profile
WHO	World Health Organization

Abstract

Background: The Department of Research Methodology, Measurement, and Data-Analysis in the Behavioral Sciences faculty of University of Twente, is currently developing a Computer Adaptive Test (CAT) to assess health-related quality of life (HRQoL) in patients with Chronic obstructive pulmonary disease (COPD).

This CAT will be based on of 3 generic item banks (derived from PROMIS) and one COPD-specific item bank (COPD-SIB). While the generic PROMIS item banks were already validated, the COPD-SIB was developed recently and its psychometric properties have yet to be evaluated. In order to contribute to the development of the CAT, this thesis aims to evaluate the psychometric properties of the COPD-SIB and (if necessary) formulate suggestions for improvement of such, so that it can be included in the final CAT item bank without worries.

Methods: The item bank was analyzed using a latent variable model. This was done in two complementary steps. Firstly, an exploratory (factor) analysis was performed to determine the number of latent variables. Secondly, a confirmatory analysis (IRT) was performed in order to assess item quality and measurement precision as a function of the latent trait.

Results: Exploratory factory analysis revealed that the item bank is reasonably unidimensional. IRT analysis showed that half of the items were sufficient discriminative. However, in 52 out of 66 items one of the categories was superfluous, or categories were not logically ordered. Test measurement was most precise around $\Theta \approx 0$.

Conclusion: Though the item bank is sufficiently unidimensional, items that were striking in exploratory as well as in confirmatory analysis should be either excluded or adjusted before being used in the CAT. Items that showed low discrimination should be rephrased. Additionally, response categories should be merged.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease that reduces airflow to the lungs and thus causes breath-related problems. Due to accelerated lung function decline, symptoms such as shortness of breath, chest tightness, coughing and a lack of energy frequently occur. One of the major causes for COPD is long-term consumption of tobacco (Decramer et al., 2012). According to the World Health Organization (2013), COPD is the third-leading cause of death, killing over 3 million people a year. While nowadays approximately 10% of the world population is affected by this disease, its prevalence and mortality is expected to further increase in the oncoming decades (Decramer et al., 2012; Lopez et al, 2006). COPD is often medicated with bronchodilators or inhaled glucocorticosteroids. Bronchodilators help to improve the airflow, while the inhaled steroids help to reduce airway inflammation. Unfortunately, such treatments are only palliative and do not lead to a cure so far (Pauwels, Buist, Calverley, Jenkins and Hurd, 2001). Hence, the goal of treatment should primarily be focused on reducing the impact of the disease on the patient's life and thus preserve the remaining health-related quality of life (HRQoL).

1.1. Health-Related Quality of Life

A review of the relevant literature indicated that the terms quality of life (QoL) and health-related quality of life (HRQoL) are often used interchangeably. However, a clear distinction should be made between the two.

According to a definition of the World Health Organization (WHO), the overall concept of QoL focuses on the individuals' perception of their position in life in relation to their goals, expectations, standards and concerns (WHOQoL Group, 1998). Consequently, one needs to consider that QoL has a fundamentally different meaning for healthy people than for people that are affected by a certain kind of disease. Therefore, it is important to distinguish between QoL in healthy and sick people. While the overall concept of QoL encompasses all aspects of life that affect an individual's experience of daily well-being (including aspects such as financial security, job satisfaction or family life), HRQoL focuses on particular aspects of QoL in relation to a certain kind of disease. The WHO not only provides a definition of QoL, but also specifies HRQoL. According to them, HRQoL is a multidimensional construct based on the subjective perspective of health encompassing physical, psychological and social functioning. The concept has evolved since the 1970s, when Boucot (1969) first emphasized the moral obligation of physicians to not only focus on

extending patients' lives, but also to provide for a certain degree of satisfaction in this extended life. Of course, the degree of satisfaction, in this connection, is a highly subjective concept which strongly depends on the patients' personal sensation of how badly the disease impacts their QoL.

Though objective somatic measurements (such as lung function or FEV1) do provide valuable information about the stage and process of COPD, they hardly offer an insight into the patients' personal perception of the disease. Not only Boucot has emphasized the importance of the role of the subjective experience of the disease. Also various other studies suggest that patients' personal perceptions are an important contributor to adequate treatment, as they often include a variety of subjective factors and aspects that are not open to "outsiders". Diagnoses, derived from only objective somatic measurements, may lead to insufficient estimation of the actual health status, as they miss the assessment of these subjective factors. In their study Koller and Lorenz (2002), for example, examined the subjective perception of their health status in patients who underwent breast-preserving surgery. Results showed that their subjective perception significantly differed from the objective estimation of the health status diagnosed by means of somatic measurement. Consequently, patient-derived data, such as HRQoL, has steadily gained acceptance as an essential element in clinical research (Miller, 2002), since it also addresses the patients' subjective perception and thus can provide greater insight into the actual condition of the patient. With this knowledge, treatments and medication can be adapted more suitably to the patient's needs, which, in turn, will optimize patient management and thus the effectiveness of therapeutic interventions.

1.2. Measuring HRQoL

There are several instruments available for measuring HRQoL. On the one hand, there are generic instruments. These instruments aim to measure universally-relevant constructs, resulting in scores that can be compared across a broad range of health problems (Beattie, Golledge, Greenhalgh, and Davies, 1997). They are also a valuable tool to generate normative data by using these instruments within healthy populations (Beattie et al., 1997) These normative data can then be used to compare them to different patient groups Unfortunately, such a broad applicability may result in limited suitability for specific patient populations, as these instruments are potentially less sensitive in detecting small but clinically important differences in treatment effects (Mehta et al., 2003). In COPD, the most commonly used generic questionnaires are the Form 36-item Questionnaire (SF-36), the Sickness Impact

Profile (SIP) and the Nottingham Health Profile (NHP) (Both, Essink-Bot, Busschbach, and Nijsten, 2007; Mueller-Buehl et al., 2003).

On the other hand, there are disease-specific instruments. These instruments specifically focus on aspects in relation to a certain disease. As an effect, they are more responsive to change and thus represent an attractive accompaniment to generic instruments (Beattie et al., 1997). Most commonly used COPD-specific instruments are the Chronic Respiratory Questionnaire (CRQ) (Guyatt, Berman, Townsend, et al., 1987), the St George's Respiratory Questionnaire (SGRQ) (Jones, Quirk, Baveystock, and Littlejohns, 1992) and the Breathing Problems Questionnaire (BPQ) (Hyland, Singh, Sodergren, and Morgan, 1998; Hyland, Bott, Singh, and Kenyon, 1994).

Often fixed-length paper questionnaires (like the ones mentioned above) face several challenges. The most prominent problem here is the large number of items needed to obtain a reliable and valid estimation of the outcome/latent trait. Unfortunately, large numbers of items mostly lead to long and tiring questionnaires. This, in turn, is less productive for measurement precision, since respondents often begin to answer the questions inadequately after some time (Herzog and Bachma, 1981). Shortening the questionnaire, on the other hand, is also not a suitable solution, as too short questionnaires often lack validity, since they might miss some important aspects. In order to find a solution to this problem, the Department of Research Methodology, Measurement, and Data-Analysis in the Behavioral Sciences Faculty of the University of Twente is currently developing a Computer Adaptive Test (CAT) to assess HRQoL in COPD patients.

1.3. Measuring HRQoL adaptively

A CAT, in contrast to fixed-length paper questionnaires, is a computer-based questionnaire which successively administers items according to a certain item selection algorithm. Each item is selected on the basis of the information gathered from the previous answered item. In the context of measuring poor HRQoL: If a test-taker, for example, answers positive to a particular item, the item displayed next will be suited to that answer and thus an item which is stronger connected to poor HRQoL will be displayed. In this manner, only relevant items are selected and greater measurement precision can be achieved. The items are selected from a collection of items, known as an item bank. The item bank, used for the CAT that is currently being developed at the University of Twente is based on three generic and one disease-specific item bank.

1.3.1. The generic item banks

The three generic item banks were selected from the Patient Reported Outcomes Measurement Information System (PROMIS) and aim to measure three crucial domains of HRQoL in COPD: fatigue, physical and social functioning. These three domains were selected based on interviews with COPD patients and healthcare professionals (Paap, Bode, Lenferink, Groen, Terwee, Ahmed, Eilayyan, and van der Palen, 2014; Paap, Bode, Lenferink, Terwee, and van der Palen, 2015).

PROMIS is a system which entails numerous self–reported health information gathered from patients by asking questions regarding their subjective perception of their physical, mental and social well–being. In this manner, PROMIS aims to provide clinicians and researchers access to efficient, precise and valid self-reported health measurements. All metrics for each domain have been developed and evaluated according to a specific set of standards. Furthermore, multiple studies were completed in order to validate the instruments. Among them, for example, a validation studies for the physical functioning scales (e.g. Jensen, Potosky, Reeve, Hahn, Cella, Fries, and Moinpour, 2015) or for the anxiety and depression symptom (e.g. Irwin, Stucky, B., Langer, Thissen, DeWitt, Lai, and DeWalt, 2010).

1.3.2. The disease-specific item bank

The disease-specific item bank was developed recently on the basis of four successive steps.

First, it was determined which topics should be covered in the item bank. Topics were identified by conducting a literature review and through analyzing interviews with patients conducted in a previous study (Paap, et. al, 2014; Paap, et al., 2015). Second, relevant items were selected from existing COPD-specific instruments, based on the findings of step 1. Third, gaps between the topics covered by the instruments and the topics found in step 1 were identified. To fill in these gaps, new items were written. Finally, cognitive interviews were conducted and items were improved based on patients' feedback. The process of item generation for the COPD-SIB is displayed in *Figure 1*.

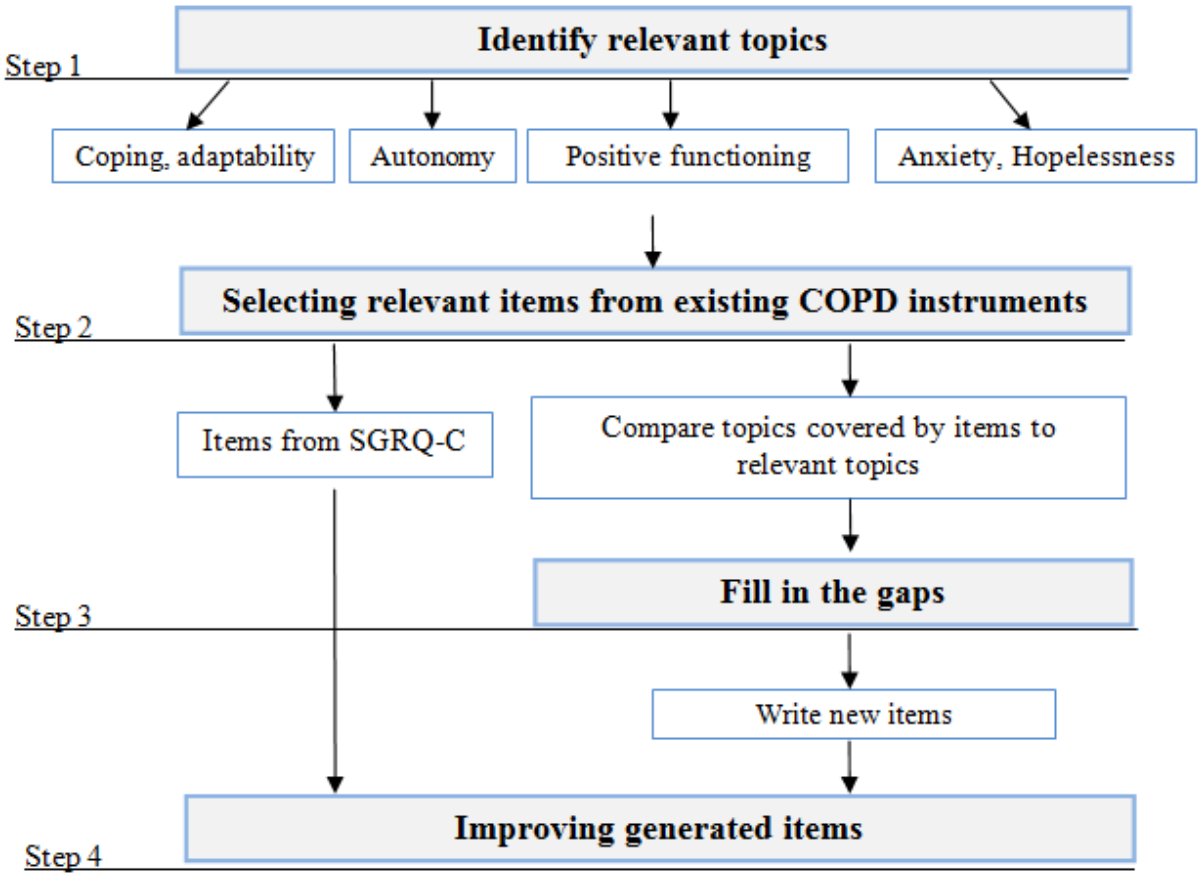

Figure 1: Schematic display of the item generation process for the COPD-specific item bank

Step 1: Identify relevant topics that should

In order to identify relevant item banks, two studies have been implemented. In these two studies, COPD Patients and health care professionals have been interviewed. First, both groups of respondents were asked to freely describe which aspects of life they find to be impacted by COPD. In the second phase, the respondents were presented 16 different HRQoL domains, gathered from PROMIS. Test-takers were asked, first to select five domains most relevant to them, and then to rank them in an order of priority. Additionally, respondents were requested to verbalize their thoughts while making their choices. Combining patient and HCP perspective the following set of PROMIS domains for assessing HRQoL in COPD were proposed: Fatigue, Physical function, Satisfaction with/ability to participate in social roles and activities, Companionship, Emotional support, Instrumental support and Depression. During the open question interview and the "think out loud" task, the respondents frequently mentioned additional other things that appear to be important to them, but were not yet (sufficiently) covered by the PROMIS item banks From these statements, several additional item themes that were not covered by PROMIS have been derived: (1) Coping with disease / symptoms, adaptability, (2) Autonomy, (3) Anxiety course / end-state of the disease,

hopelessness, (4) positive psychological functioning (5) situations triggering or enhancing breathing problems (6) symptoms (7) activity (8) impacts

Step 2: Selecting relevant items from existing COPD-specific instruments

In step two, the relevant literature was reviewed with the objective of investigating which disease-specific questionnaires are most commonly used in COPD. The St. George Respiratory Questionnaire for COPD patients (SGRQ-C) was taken as a starting point here, since it is a widely used tool to asses HRQoL in COPD patients and contains many items of good quality (Paap, Brouwer, Glas, Monninkhof, Forstreuter, Pieterse, and van der Palen, 2015). Items from the SGRQ-C that did not show too much overlap with the previously selected PROMIS domains Fatigue, Physical function, Satisfaction with/ability to participate in social roles and activities, Companionship, Emotional support, Instrumental support and Depression were included in the initial COPD-SIB. Subsequently, other HRQoL questionnaires commonly used with COPD patients were identified, and relevant items from those questionnaires were selected as well. The questionnaires were: the Quality of Life for Respiratory Illness Questionnaire (QoLRIQ), the COPD specific HRQoL Questionnaire (VQ11) and the 26-item Maugeri Respiratory Failure Questionnaire (MRF-26). Inclusion criteria were: a) the items did not show too much overlap with already selected SGRQ-C items and PROMIS items that were to be included in the CAT; b) they pertained to the three themes found in step 1; and c) permission from the developers of the questionnaire for use of these items.

Step 3: Fill in the gaps

After items had been selected from existing instruments, the topics covered by these items were compared to the themes frequently mentioned in the patient interviews (cf. step 1). Gaps were identified and new items were written on the basis of the themes (if possible, patient quotes were used for item generation).

Step 4: Improving generated items

In order to evaluate the item content and improve the item wording, the generated items then underwent a series of adaptations. Due to practical reasons the SGRQ-C items and selected items from other existing COPD-specific instruments as well as newly written items were tested in two parallel interview rounds, both using the Three Step Test Interview method (see Hak, Van der Veer, Jansen (2004) for further explanation). SGRQ-C items were presented to

20 COPD patients at the MST department ‘pulmonary medicine’. Thirteen of the respondents were female and seven were male. The mean age was 63.25 years (SD=11.37). The other items were presented to 16 respondents, whereof 56% were recruited through a hospital in Enschede, 31% through a hospital in Zwolle, 6% through a hospital Eindhoven and 6% through a hospital Meppel. Ten respondents were male and six were female. Mean age was 72.19 years (SD = 5.75). In the interview rounds, the statements of the respondents were evaluated iteratively and the items were adjusted according to the information gathered. This thesis focuses on the final version of the COPD-SIB (see *Appendix A)*, since this is the version to be used in the CAT.

1.4. Purpose of thesis

According to Embretson und Reise (2000), a CAT can only be as good as the item bank it is based on. Especially in CATs, high demands are put on the given item bank, as adaptive reduction of the test length through elimination of "inferior" items, can very much affect the course of the test. While the generic PROMIS item banks were already validated in the USA, the COPD-SIB is currently being developed and its psychometric properties have yet to be evaluated. In order to pave the way to the development of the CAT, this thesis aims to evaluate the psychometric properties of the COPD-SIB. Additionally, suggestions for improvement of the item bank will be formulated on the basis of the findings, so that it can be included in the final CAT item bank without any difficulty or reservations. The items for the COPD-SIB were selected and designed with the expectation that they tap into a single construct, while covering all relevant themes to ensure content validity. This thesis therefore addresses the following research question:

- What is the dimensional structure of the COPD-specific item bank?

When the item bank was designed, the developers operated under the assumption that it would measure a unidimensional construct, namely HRQoL. However, items from a wide range of themes were included to ensure content validity. Considering the process of item generation it can be expected that the COPD-SIB consists of the following eight subdomains[1][2]:

(1) Coping with disease/symptoms, adaptability

(2) Autonomy

(3) Anxiety about the course/end-state of the disease, hopelessness

[1] Domains 1-5 are derived from the think aloud and open question interview (cf. step 1 of the item generation process).

[2] Domains 6-8 are derived from the SGRQ-C.

(4) Positive psychological functioning

(5) Situations triggering or enhancing breathing problems

(6) Symptoms

(7) Activity

(8) Impacts

1.5. Hypothesis

Derived from the above mentioned research question and the given assumption of six subdomains, the following hypothesis was formulated:

- The COPD-specific item bank has a multi-factor structure (is multidimensional).

1.6. Research strategy

To test this hypothesis the item bank will be analyzed using a latent variable model. This will be done in two steps. In the first place an exploratory (factor) analysis is performed to determine the number of latent variables. Secondly a confirmatory analysis (IRT) will be performed in order to assess item quality and measurement precision as a function of the latent trait.

2. Methods

2.1. Study design

Since the CAT will consist of both the COPD-SIB as well as the three generic PROMIS item banks, an overall questionnaire was developed to be able to evaluate all four item banks. A feasibility study revealed that an amount of 100 items per questionnaire is appropriate. Due to the fact that including all 4 item banks in one questionnaire would lead to an infeasible amount of items, this overall questionnaire was divided over three test versions (so-called ”booklets“), with each booklet including a certain number of items from each item bank. Since, the purpose of this thesis was to analyze the dimensional structure of the COPD-SIB, only items stemming from this item bank were included in the analysis. The COPD-SIB items were divided over the three booklets according to an Anchor-Test design (Sinharay and Holland, 2006), where particular items were systematically included in all 3 versions. Through the use of anchor items the three booklets can be merged again for later confirmative analysis. Figure 2 illustrates an overview of the items per booklet.

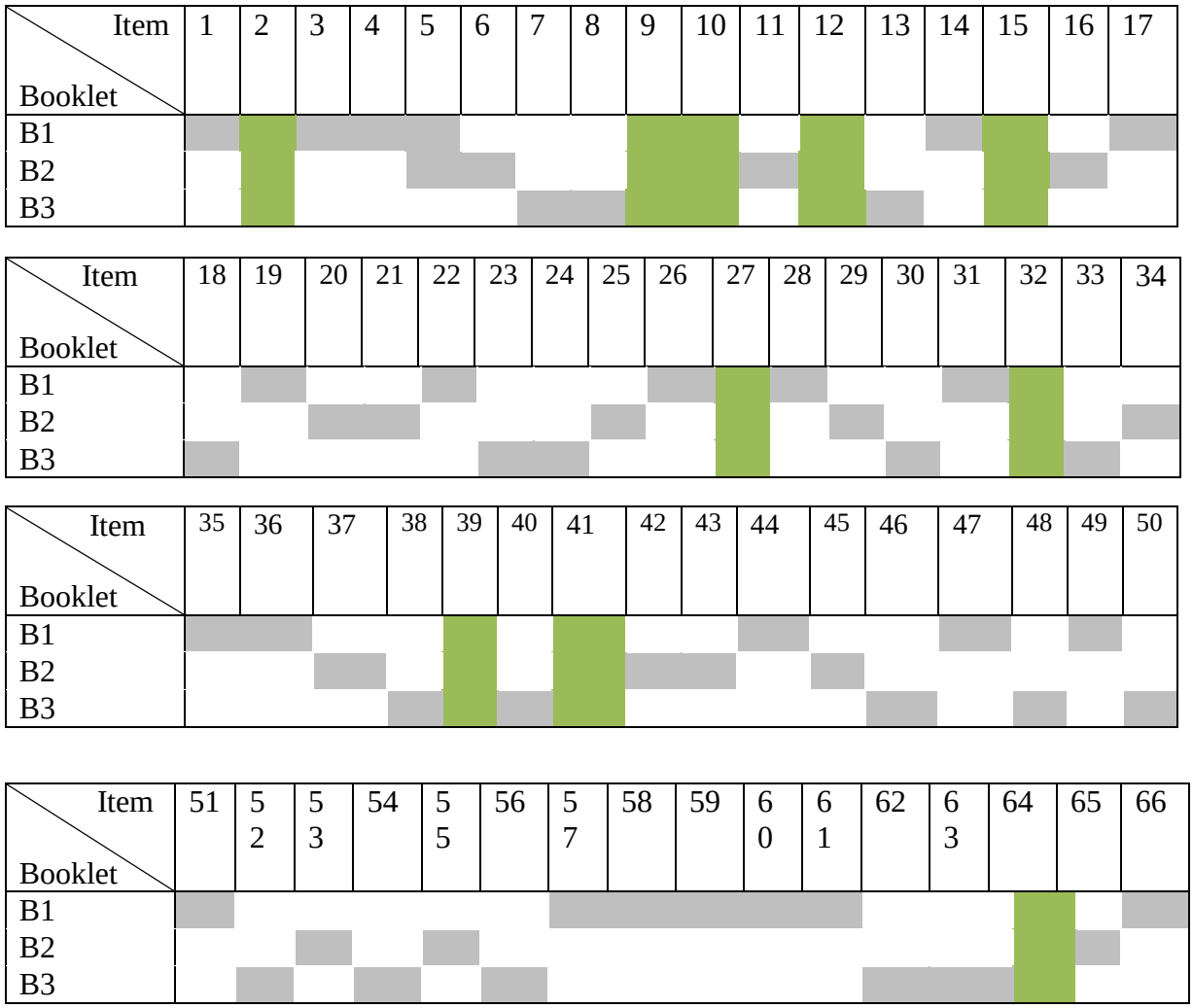

Figure 2. Overview items per booklet

2.2. Instrument: the COPD-specific item bank

The final COPD-specific item bank consists of 66 items, all scored on a 5-point Likert scale. As some items stem from existing COPD questionnaires (cf. section 1.3.2.), several adjustments were necessary to generate a coherent questionnaire. Items 1 to 8 stem from the QoLRIQ and tap into the theme: situations that provoke or worsen respiratory problems (Maille, Koning, Zwinderman, Willems, Dijkman, and Kaptein, 1997). Originally, those items are scored by means of a 7-point Likert scale. After adaptation, they were scored on a scale from 1= ”not at all“ (Dutch: helemaal niet) to 5= ”very strongly” (Dutch: heel erg). Additionally, the original recall period of 4 weeks was removed. The items were translated from English to Dutch. Item 29 also stems from the QoLRIQ and was labeled into the theme autonomy. Unlike the other items stemming from this questionnaire, this item was rated on a 5-point Likert scale ranging from 1 = strongly disagree (Dutch: zeer mee oneens) to 5 = strongly agree (Dutch: zeer mee eens), since this fits more into the context of all other items tapping into this theme.

Item 9 stems from the VQ11 and taps into the theme “functional status“(Ninot, Soyez, and Préfaut, 2013). The item was first translated from English to Dutch and then rephrased from ”I feel unable to achieve my objectives“ to “Because of my COPD, I feel unable to achieve my objectives“ since adding “because of my COPD…” was more consistent with other items. Answer options were changed into a 5-point Likert scale, ranging from 1 = strongly disagree (Dutch: zeer mee oneens) to 5 = strongly agree (Dutch: zeer mee eens). Items 34 to 48 and items 57 to 65 stem from the SGRQ-C and are tap into the sub-themes of ”impact“ (items 34 to 45 and 57 to 60) and ”activity“ (items 46 to 48 and 61 to 65) respectively (Meguro, Barley, Spencer, Jones, 2007). No translation was needed, since there is an official Dutch version of the SGRQ-C available. Originally, these items were scored on a dichotomy true/false scale. In order to generate a coherent questionnaire, the scoring was likewise changed to a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree. Items 57 to 65 are also scored by means of a 5-point Likert scale. However, 1 here equals”never“ (Dutch: nooit), 2 means “seldom” (Dutch: zelden), 3 indicates ”sometimes“ (Dutch: soms), 4 means ”often“ (Dutch: vaak) and 5 equals ”always“ (Dutch: altijd). Item 40 was rephrased from “I get exhausted easily“ to ”I get tired easily“. Item 42 was rephrased from ” My chest trouble is a nuisance to my family, friends or neighbours“ to “I feel that my chest trouble is a nuisance to my environment (e.g. family, friends or neighbours)”. Item 64 was rephrased from “Walking

outside on the level" to "Going for a walk". In the original version of the SGRQ-C, items 48 and 49 form one composite question. However, the developers of the item bank decided to split up this question and generate 2 separate items, based on patient feedback (Paap, et al., 2014). Item 30 and 32 to 34 stem from the MRF-26. The original version of the MRF-26 assumes unidimensionality and, hence, aims to measure one overall theme, namely HRQoL (Vidotto, Carone, Jones, Salini, and Bertolotti, 2007). However the domains given in the SGRQ-C fit quite well to these items. Item 30 therefore was assigned to measure "impact", while item 32 to 34 measure "activity". The items were translated from English to Dutch and are scored on the same 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree.

Item 10 to 28 and 31 were self written and are related to the four themes (1) coping with disease/symptoms, adaptability, (2) autonomy, (3) anxiety course/end-state of the disease and (4) positive psychological functioning. Those items are scored by means of a 5-point Likert scale, ranging from 1 = strongly disagree (Dutch: zeer mee oneens) to 5 = strongly agree (Dutch: zeer mee eens). Item 50 to 57 are also self written and scored on a 5-point Likert scale. However, 1 here equals "never" (Dutch: nooit), 2 means "rarely" (Dutch: zelden), 3 indicates "sometimes" (Dutch: soms), 4 means "often" (Dutch: vaak) and 5 equals "always" (Dutch: altijd). Item 15, 19, 20, 22, 23, 25, 27, 49 and 51 to 56 tap into theme 1. Item 12, 13 and 17 tap into theme 2. Item 10, 11, 14, 16, 21, 24, 26 and 50 tap into theme 3.
An overview of the scoring, the original items and how they were translated is represented in *Appendix B.*

2.3. Data collection

The data was collected by sending the questionnaires to several hospitals and clinics in the Netherlands with the request to hand out the questionnaires to COPD patients. Completed questionnaires were received from the CW Hospital in Nijmegen, the Medisch Spectrum Twente in Enschede, the Scheperziekenhuis in Emmen, the Expertisecentrum voor chronisch orgaanfalen (CIRO) in Horn, the St. Lucas Andreas Hospital in Amsterdam, the Martini Hospital in Groningen and from several general practitioners and physiotherapists based in the area of Twente (Overijssel province).

2.4. Data analysis

In order to investigate the dimensionality of the item bank, it was chosen to analyze the COPD-SIB with the help of two complementary statistical methods: exploratory factor

analysis and confirmatory IRT analysis. In exploratory analysis, each booklet was analyzed separately, as merging the data would generate too many missing values and exploratory analysis cannot deal with such a large number of missings. Thereafter, the three data files were merged and an IRT analysis was conducted on these merged files. Here, the decision was taken to merge the files because (a) IRT is able to deal with such a huge amount of missing data and (b) in this manner a better picture of all items acting together is provided.

2.4.1. Exploratory analysis

To test the assumption made by the developers, exploratory factor analysis was performed using the ”Statistical Package for Social Science“ (SPSS), version 20. In order to execute an exploratory data analysis, it is necessary to have a complete dataset. Therefore, it was chosen to first generate a complete dataset by implementing two-way imputation to fill in missing data.

2.4.2. Two-way imputation

Two-way imputation is a method for imputing missing data that takes into account both, person effects as well as item effects (van Ginkel, and van der Ark, 2010). For a detailed explanation of how each score for each missing value is computed the reader is referred to van Ginkel and van der Ark (2010). In order to execute this imputation, the data requires certain preparation.

A requirement of two-way imputation is that there are less than 5% overall missing values, since imputing more than 5% of missing values would distort the picture given by the dataset (van Ginkel, and van der Ark, 2010). This overall rate of missing values is composed of two complementary types of missings: items that show a lot of missing values (hence, column-wise) and persons, who systematically did not respond to a conspicuous number of items (hence, row-wise). Therefore, the first step of data analysis was hence, to examine the dataset for these two types of missing data. Column-wise “missings” (“missings per item”) were examined by executing a frequency analysis for each item. Items which showed more than 20% missing values were considered to show systematic missings or missings not at random (MNAR), as it is reasonable to assume that the data is missing for a specific reason when there is such a striking number of non-responses. Row-wise missings (“missings per person”) were rectified by calculating a new variable (Nmiss) and manually deleting persons who did not respond to more than 60% of the items. This criterion was derived from a recommendation of van Ginkel, and van der Ark, (2010). Items and persons who showed

conspicuous missing data rates were deleted manually and, thus, excluded from further analysis. As a complement to this, an analysis of patterns of missing values was executed in order to reveal what percentage of missing values is left. When the percentage of missing values was less than 5% of all total cells (items x persons), two-way imputation was possible. Another necessary preparation was to recode contra-indicatively worded items prior to the imputation. This was because two-way imputation assumes that a higher item score is indicative of a higher score on the construct that is being measured (van Ginkel, and van der Ark, 2010). After the dataset was accurately prepared, two-way imputation was implemented. Now that the requirement of a complete dataset was given, exploratory factor analysis could be executed.

2.4.3. Exploratory factor analysis

First, answering categories with less than 10 observations per category were merged. In IRT each item is estimated with m-1 thresholds (m = number of answering options). In order to estimate these parameters accurately, sufficient observations for each answering option are necessary. Therefore, answering options were systematically merged first, as a preparation for IRT analysis and also to provide better comparability between results from exploratory factor analysis and confirmatory IRT analysis. Subsequently, an inter-item-correlation matrix was reviewed to detect whether there are items that show very weak or negative correlations with other items. These items were assumed to not fit into the unidimensional model and hence gained extra attention when evaluation the factor analysis.

Next, exploratory factor analysis (EFA) was performed, in order to determine the number of latent factors within the COPD-SIB (cf. hypothesis). Maximum Likelihood estimation was used to calculate a multi-factor solution in the first place. The results of this multi-factor solution were evaluated by examining the generated scree plot. This examination technique was interpreted according to certain a recommendation of Fayers and Machin (2000). According to them, an interpretation of a scree plot is subjective, but the most common rule of the thumb applied to interpret this plot is to focus on the change in slope of the curve. For example: There are two factors scoring above 1. Then, a change in slope occurs and all later factors form a distinct accurate sloping line which slowly moves towards 0. The 2 factors before this change in slope can be interpreted as evidence of a number of factors.

The suggested factor solution resulting from these two examinations was then further examined by executing a second factor analysis with a reduced number of factors to be extracted. The magnitude of the resulting item loadings was explored in order to interpret the

prevailing factors. A factor can be interpreted as prevailing if a cluster of items can be ascribed striking obviously to load on one certain factor (Embretson and Reise (2000), This means that an item can be ascribed to one particular factor when item loadings on this particular factor are higher than the loadings on a second or thirds factor. Items that were conspicuous were not excluded for further IRT analysis so far, since it was interesting to compare findings from exploratory analysis with those from IRT analysis.

2.4.4. IRT analysis

IRT analysis was conducted using R-statistics. IRT analysis was applied using the R package ltm. The whole syntax used is displayed in *Appendix C*.

The principal concept in IRT is the Item Characteristic Curve (ICC), which is a graphical representation of the probability a person has for choosing a specific category depending on the latent trait (Fayers and Machin, 2000). The latent trait is called theta (Ɵ) in IRT. Theta is a standardized estimate of a ”true“ score for the latent trait. In our case, theta would indicate the standardized value of the respondents' perception of HRQoL.

There are 3 types of IRT models that can be distinguished: 1PL, 2PL and 3PL. The models are distinguished according to the number of parameters they entail. A 1 PL model entails only one parameter, the so called “b” parameter, which reflects the positioning of an item on the latent trait. This parameter is also called *difficulty* parameter as its position on the latent trait scale indicates how ”difficult” an item is, depending on the magnitude of the latent trait. Hence, if an item has low *difficulty* it is easier to answer positive (or right) to that item and the item would thus measure a lower magnitude of the latent trait. Translated to this thesis this means: if a respondent experiences a high level of discomfort, s/he is more likely to answer a "difficult" item in a positive way. 2PL models (as the name already suggests) additionally include a second parameter, the so called „a” parameter. This parameter reflects the steepness or slope of the ICC. This parameter is also called *discrimination* parameter, as it reflects the ability of an item to discriminate between high and low thetas. This parameter is very important, as it also determines the amount of information (measurement precision) provided by an item. Items with higher discrimination parameters provide more information and vice versa (DeMars, 2010). In this thesis, items with a discrimination parameter > 0.8 were determined as appropriate. This criterion was derived from recommendation of Walter, Becker, Fliege, Bjorner, Kosinski, Walter, and Rose (2005), who argued that items with a discrimination parameter < 0.8 are likely to be finally included in a CAT in only 0.05% of all cases.

A third IRT model is the 3PL model. This model includes a third parameter, which is the guessing parameter. This parameter is more important in educational testing, as it models the probability of answering an item ,"correctly" by chance. For this thesis, this parameter is not important, since the COPD-SIB is designed to estimate the subjective construct of COPD-related discomfort, where there is no "wrong" or "right". The probability to guess correctly is hence not given. Although it is advisable to use 1PL models, when there are sample size constraints (Yu, 2013), in this thesis a 2PL model was used, since it is important (regarding the development of a CAT) to examine the discrimination ability for each item.

Another statistical characteristic of IRT is information. Information can be compared to concepts like reliability and measurement precision. Information can be calculated both at item as well as test level, and is usually evaluated by inspection of the Information Function. The Item Information Function (IIF) shows how much information each item provides for different theta values. By summing the item information, test information can be calculated. In this way, the IRT analysis is able to provide an insight in for which of the theta values the test provides the most accurate measurement (Embretson and Reise, 2000). In this thesis, the test information is of crucial importance, as it can tell us whether the COPD-SIB gives rich information only for a small range of theta or not. Regarding a graphical presentation of the test information, the test information curve should be as broad and as high as possible. If this is the case, the item bank would indicate good information (height) for a wide range of theta (broadness).

In this thesis the items were analyzed using the generalized partial credit model (GPCM) for polytomous items. This model allows for a variable pitch of the various curves of the individual item response categories (Muraki, 1992). It involves (in addition to the usual a and b parameters) a second, from the b parameter originated, b1 parameter. This b1 parameter is also called *item threshold* parameter and specifies the location of all response categories of all items on the latent trait. Graphically, this parameter can be located where two adjacent category response curves intersect. This point thus indicates "the point on the latent-trait scale where one category response becomes more likely than the preceding response" (Embertson and Reise, 2000, p.111). As IRT can handle missing data, the data of all three booklets were merged and all datasets were analyzed together in this analysis. Firstly, a 2PL GPCM model was calculated. Item parameters for each item were reviewed in order to get a first overview of conspicuous items. Items were labeled as "unsatisfactory" when (1) they could not discriminate between the different theta values (low a parameter) and (2) their b parameters

had a value higher than 3 or lower than -3. Next, Item Information Curves for each item were reviewed. Lastly, the overall test information curve for each booklet was examined.

2.5. Respondents

Since each booklet was analyzed separately in exploratory data analysis, there are three different groups of respondents.

In Booklet 1, 108 respondents answered to the items, of whom 54.3% were male, 42.4% were female and 3.3% did not indicate their gender. Mean age was 67.7 years (*SD*= 8.40).

In Booklet 2, 110 respondents answered to the items, of whom 51% were male, 45.9% were female and 3.1% did not indicate their gender. Mean age was 67.8 years (*SD*= 8.81).

In Booklet 3, 154 respondents answered the items, of whom 48.9% were male and 49.6% female. 1.4% of the respondents did not answer this item. Mean age was 66.0 years (*SD*= 9.84). Inclusion criteria were a medical diagnosis of COPD, adequate oral, reading and writing mastery of the Dutch language and being able to complete a questionnaire.

In the IRT analysis, where the files were merged the demographics were as follows: in total 372 respondents responded to the items. 51.08% were male, while 47.75% were female and 1.17 % did not answer this item. Mean age was 67.2 years (*SD* = 9.01).

Table 1

Demographics

analysis	N	gender		age	
		male	female	mean	SD
booklet 1	108	54.3%	42.4%	67.7	8.40
booklet 2	110	51%	45.9%	67.8	8.81
booklet 3	154	48.9%	49.6%	66.0	9.84
merged files	372	51.08%	47.75%	67.2	9.01

3. Results

3.1. Exploratory analysis

Overall, all three item banks showed less than 5% missing values and were, hence, appropriate for two-way imputation. However, Booklet 2 and Booklet 3 both included items with more than 20% missing rates (Item 6 in B2 and Item 7 and 8 in B3). All these items were more related to asthmatic problems instead of only COPD and, hence, excluded from further analysis. Booklet 2 and 3 also contained persons with more than 50% missings. In B2, two persons and in B3 three persons were likewise excluded from further analysis. Inspection of inter-item correlations indicated that several items correlated negatively with other items in each booklet. Item 10, which is an anchor item, was striking since it correlated negatively with other items in all three booklets (19 negative correlations in B1, three negative correlations in B2 as well as in B3). In all three booklets item 10 did not load on the first factor (B1: $\lambda < .400$; B2: $\lambda < .400$) but rather loaded on a second factor (B3: $\lambda = .659$). Scree plots for each item bank were strongly suggestive of a single factor.

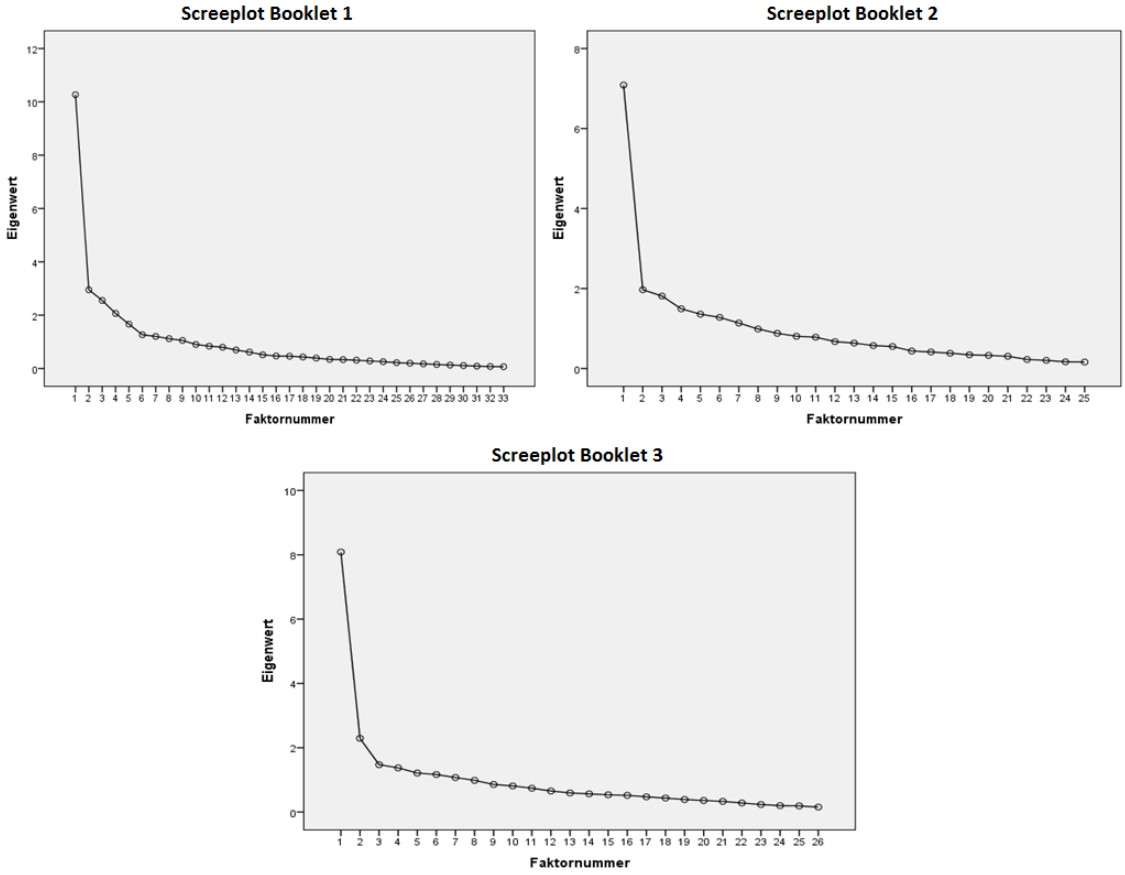

Figure 3. Screeplot per booklet

3.1.1. Booklet 1

Missing value analysis showed that no items or respondents had missing rates above the specified threshold (item <20%; respondents < 50% missing values). Analysis of patterns of missing values then indicated that there were 2.363% missing values in the overall item bank. Hence, two-way imputation was possible. Inter-item correlation showed that items 10 had 19 negative correlations or correlated only very slightly with the other items. Results of factors analysis also showed that item 10 did not load on the first factor ($\lambda < |.400|$). Furthermore, factor analysis revealed that 27 out of 33 items loaded on the first factor ($\lambda > |.400|$). Items that loaded on both factors were ascribed to the factor on which the loading was higher.

Table 2

Results of Factor Analyses Booklet 1 (loadings)

Item	Factor		Item	Factor	
	F1	F2		F1	F2
1	.68		32	.520	.217
2	.562		35	.577	-.266
3	**.625**		36	**.641**	-.407
4	.597		39	**.666**	-.392
5	.583		41	.497	
9	**.606**		44	**.627**	
10			47	**.633**	.220
12	.519		49	**.609**	.275
14	**.608**	.384	58	.431	-.473
15	.284	.254	57	.579	
17	.527	.343	59	.439	-.570
19			60	.591	
22	.486	.302	61	.514	-.363
26	.315	.201	51	.557	
27	.591		64		
28	**.610**	.382	66		
31	.406				

To indicate which items showed best discrimination on the general factor, loadings that are higher than .600 are printed in bold. Items loading lower than 0.300 are omitted.

3.1.2. Booklet 2

Results of missing value analysis showed that item 6 ("being outside during the polling season") had over 21.8% missings. Since this item is more related to asthmatic problems instead of only COPD, it was decided to exclude the item from further analysis. Respondent 36 had 88.0% missings and respondent 43 had 64.0% missings. Hence, these two respondents were also excluded for further analysis. Analysis of patterns of missing value then indicated that the item bank had 1.2% missing values left. Hence, two-way imputation was possible. Inter-item correlation showed that item 16 had 17 negative correlations and item 53 had 15 negative correlations. Factor analysis revealed that those items also did not load on the first factor ($\lambda < |.400|$). However, 16 out of 25 items loaded on one factor ($\lambda > |.400|$). Items that loaded on both factors were ascribed to the factor on which the loading was higher.

Table 3

Results of the factor analyses Booklet 2 (loadings)

Item	Factor		Item	Factor	
	F1	F2		F1	F2
2	.311		34	.534	
5	.349		32	.551	
9	.484		37	.446	
10			39	.316	
12			41	**.679**	-.483
11	.381		42	**.721**	-.354
15	.453	.387	43	**.714**	
16			45	**.615**	
20	.552		53		-.336
21	**.637**		55	**.708**	
25	.590	.371	64	.305	

27	**.737**	65
29	**.711**	

To indicate which items showed best discrimination on the general factor, loadings that are higher than .600 are printed in bold. Items loading lower than 0.300 are omitted.

3.1.3. Booklet 3

Results of a missing value analysis showed that item 7 and 8 had over 20.0% missings. Item 7 had 25.8% missings, while item 8 had 23.2% missings. Since those items are more related to asthmatic problems instead of only COPD it was decided to exclude those two items from further analysis. Calculation of a new Nmiss variable showed that respondent 28 and 112 did not answer 73.0% of the items. Insufficient items and respondents were deleted and, thus, excluded from further analysis. Analysis of patterns of missing value indicated that the item bank had 2.5% overall missing values left. Hence, an imputation was possible now. Inter-item correlations showed that item 13 had 18 negative correlations or correlated only very slightly with the other items. Factor analysis revealed that 19 out of 26 items loaded on one factor ($\lambda >$ .400). Item 13 here also did not load on the first factor ($\lambda =$ -.158) but loaded on a second factor ($\lambda =$ .320). Item 10 also loaded on a second factor ($\lambda =$ .659). Items that loaded on both factors were ascribed to the factor on which the loading was higher.

Table 4

Results of the factor analyses Booklet 3 (loadings)

Item	Factor		Item	Factor	
	F1	F2		F1	F2
2	.526		38	.493	
9	.599	-.431	39	.548	
10		**.659**	40	**.625**	
12			41	.319	.322
13		.320	46	**.660**	
15	.399	-.310	48	.581	
18	**.609**		50	**.703**	
23	**.701**	-.303	52	.480	
24	.312		54		
27	.540		56	.435	.330

30	.571	62	**.670**
33	**.660**	63	**.710**
32	**.699**	64	**.612**

Note. To indicate which items showed best discrimination on the general factor, loadings that are higher than 0.60 are printed in bold. Items loading lower than 0.300 are omitted.

3.2. IRT analysis

As mentioned in section 2.5.2., the threshold value for a good discrimination parameter was a > 0.800. Overall, 33 items met this criterion, with item 29, 40 and 56 being the highest (a > 1.700), followed by item 23 with a = 1.664. The discrimination parameters of the items 19, 54, 55, 57 and 67 were not assessable. Likewise, these items showed wide ranges of b parameters. Additionally, items 10, 13, 16 and 66 indicated very poor a parameters (a < 0.300) and also had very widely ranged b parameters.

Table 5

Item Parameters

Item	Category Threshold				Dscrmn.
	Catgr.1	Catgr.2	Catgr.3	Catgr.4	
1	-0.026	0.039			0.674
2	-1.777	-0.558	-0.107	3.104	0.669
3	-0.704	-0.527	1.425		**1.160**
4	-1.414	-1.286	-0.574	1.782	**1.021**
5	-2.034	-1.082	-0.265	1.562	0.697
6	-0.838	1.994	1.195		0.397
7	1.704	0.090	1.303		0.604
8	1.963	0.849	0.466		0.495
9	-2.051	-0.792	-2.226	1.149	**0.857**
10	-11.006	4.933	1.260		0.206
11	-0.431	-1.687	2.201		0.510
12	-4.499	0.361	-5.601	2.958	0.307

13	-0.373	-3.712	4.709		0.237
14	-1.638	0.148	0.242		**1.038**
15	-3.910	-0.367	-1.092	3.323	0.546
16	-0.999	-2.688			0.286
17	-0.562	-2.280	1.002		0.695
18	-1.064	0.143	1.055		**0.919**

Table 4
(continued)

Item	Category Threshold				Dscrmn
	Catgr.1	Catgr.2	Catgr.3	Catgr.4	
19	-1672.23	-6616.29	6465.266		0
20	-0.321	-0.185	2.260		**0.871**
21	-0.714	0.211	1.012		**1.142**
22	-2.142	1.133	-0.857		0.656
23	-0.669	-1.164	1.001		**1.664**
24	-3.612	2.804	-0.883		0.317
25	-0.066	-0.995	2.216		**0.990**
26	-3.169	-0.553	-1.729		0.352
27	-1.617	0.188	0.134	2.253	**0.981**
28	-0.783	-1.157	1.132		**0.993**
29	0.228	-0.140			**1.751**
30	-1.612	0.322	-0.273	2.678	**0.881**
31	-0.698	1.120			0.661
32	-1.184	0.461	-0.207	2.418	**0.943**
33	-1.37	0.561	-0.279	2.174	**0.895**
34	-1.346	0.820	0.815		**0.874**
35	-1.207	0.731	1.226		**0.816**
36	-0.986	-1.012	-0.844	2.2	0.757

37	0.38	-0.907			0.716
39	-1.617	-0.043	-0.197	3.551	0.710
38	0.053	-1.733	2.544		0.707
40	-1.369	0.941			**1.719**
41	-2.686	0.999	0.154	3.492	0.582
42	0.140	0.176			**1.429**

Table 4
(continued)

Item	Category Threshold				Dscrmn
	Catgr.1	Catgr.2	Catgr.3	Catgr.4	
43	-1.561	0.009	-0.122		**1.428**
45	-1.843	0.622	0.922		**1.196**
46	-1.522	-0.173	-0.667		**1.057**
47	-0.143	0.114	1.988		**1.326**
48	-2.34	-0.594	-1.079	1.157	**1.312**
49	-2.119	-0.558	-2.146	0.511	**0.909**
50	-0.314	-1.374	0.616		**1.072**
51	-1.128	-0.965	0.366	2.13	**1.300**
52	0.875	0.839			**1.288**
53	0.372	0.54	1.881		0.608
54	-853.153	-2963.12	1035.843		0
55	-461,619	-8942,56	-970,198		0
56	-0.302	0.511			**1.787**
57	-10398	-3172.81	6504.197		0
58	-0.632	0.833			**1.244**
59	-1.598	-2.741	1.431	2.477	0.396
60	-1.113	-1.125	2.158	0.797	0.328
61	-1.928	0.135	1.387		**1.184**

62	-0.388	-1.176	0.578		0.547
63	-0.303	-0.861	0.321	1.191	**0.963**
64	-0.559	-0.354	1.256	1.863	**1.228**
65	-1.446	-2.77	0.475	1.543	0.398
66	-1.899	-2.728	-0.143	-1.144	0.205
67	-4099.83	-2645.669	605.209	2304.278	0

Note. To indicate which items showed best discrimination, parameters higher than 0.800 are printed in bold.

However, considering the ICC of each item, it was striking that 52 out of 66 items indicated that one of the categories was superfluous, or that categories were not logically ordered. Naturally, items which had very low discrimination parameters (as listed above) also showed category response curiosities. However, also items that actually had good discrimination performed badly in the response accuracy category. In item 29 for example (which is one of the item with best the discrimination parameter) response category 2 was superfluous.

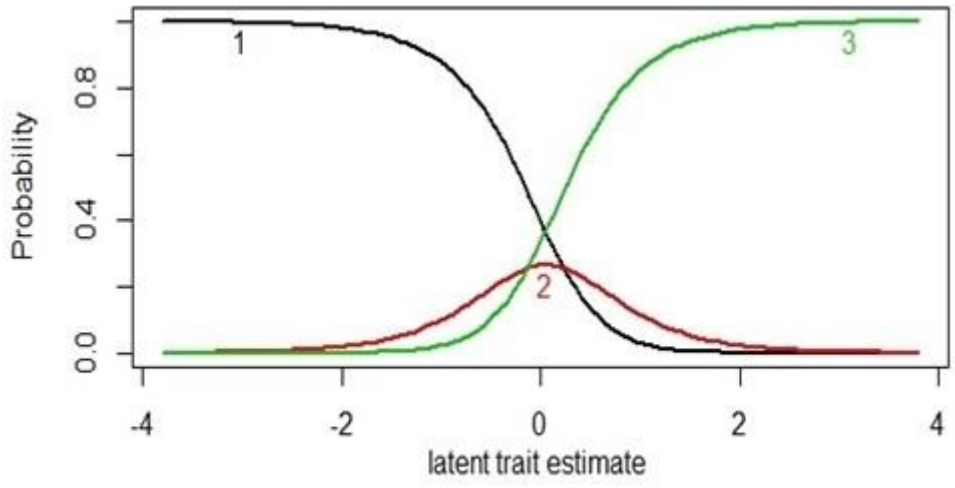

Figure 4. Item Response Characteristic Curve - Item 29

In item 23, this was also the case. Although this item had a very high discrimination parameter, response category 2 did not add value to this item.

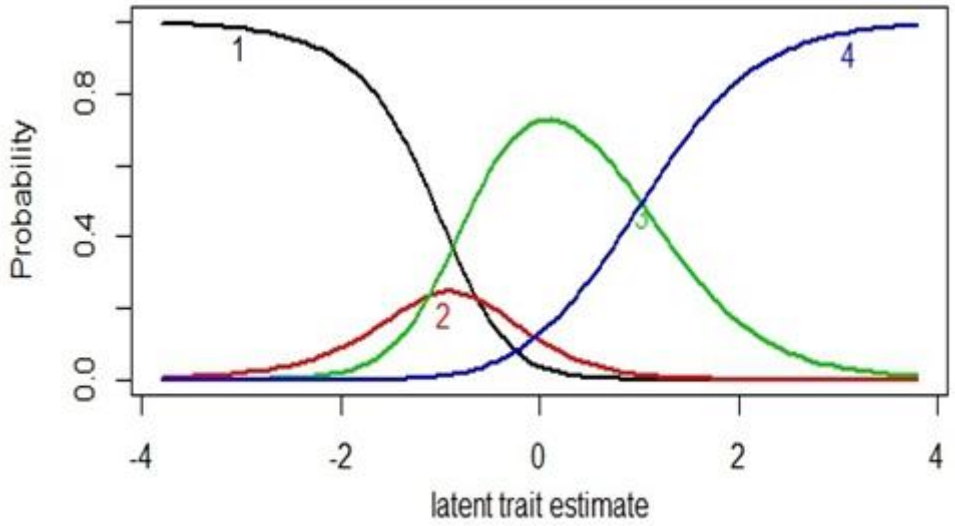

Figure 5. Item Response Characteristic Curve - Item 23

Further examples are items 42 and 43, which also had high discrimination parameters (item 42: a= 1.429; item 43: a= 1.426) but superfluous answering options.

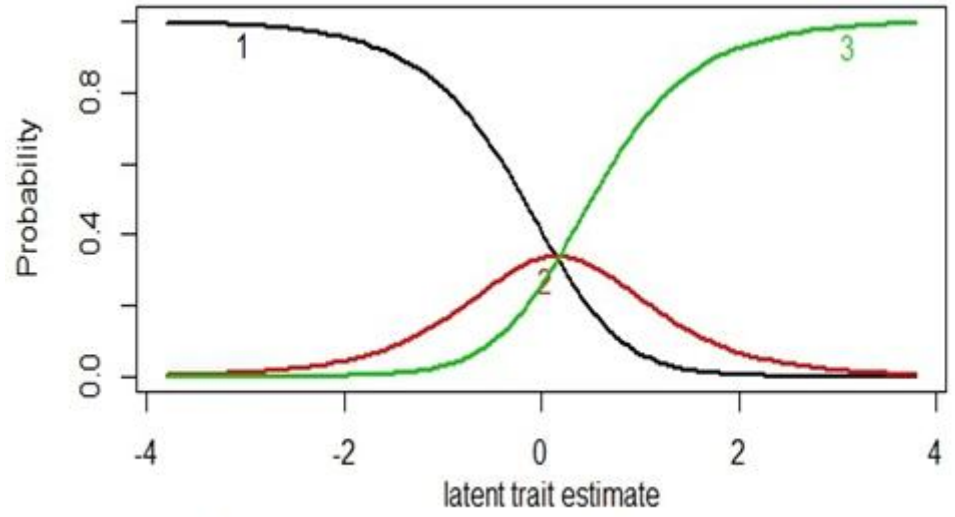

Figure 6. Item Response Characteristic Curve - Item 42

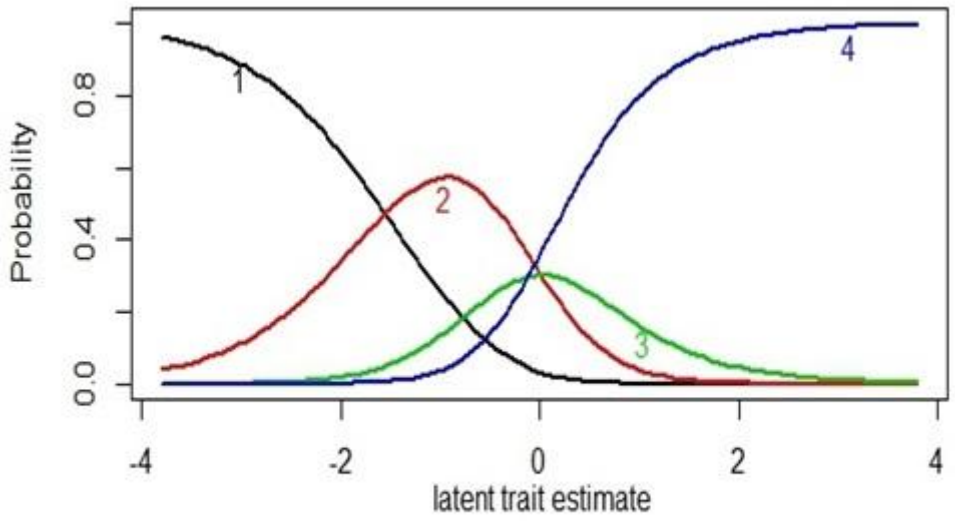

Figure 7. Item Response Characteristic Curve - Item 43

Regarding item information, results show that measurements of all items are most precise for theta values from Θ ≈ -1,5 to Θ ≈ 1. However, gaps for certain theta values can be detected. Measurement precision was low for Θ < -2 and Θ > 1.9.

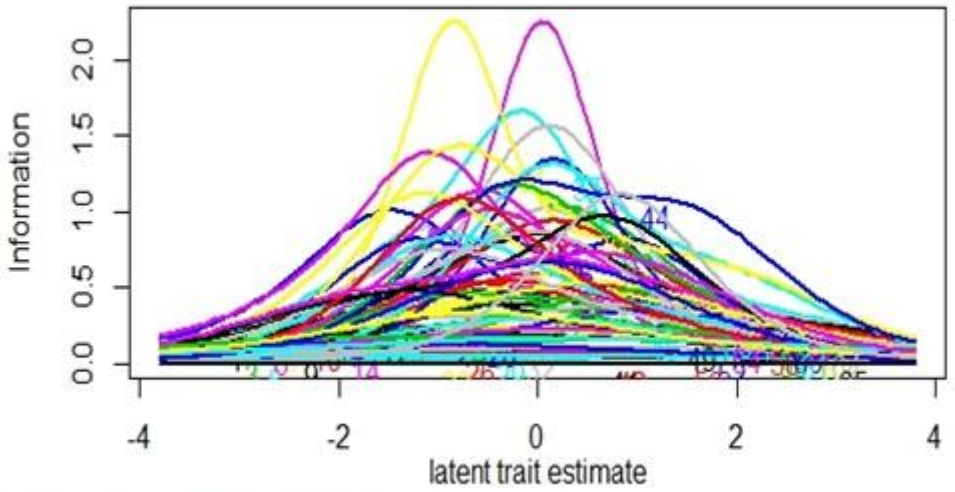

Figure 8. Item Information Curves

The test information curve was quite "peaky". The most precise measurement was given at Θ ≈ 0.

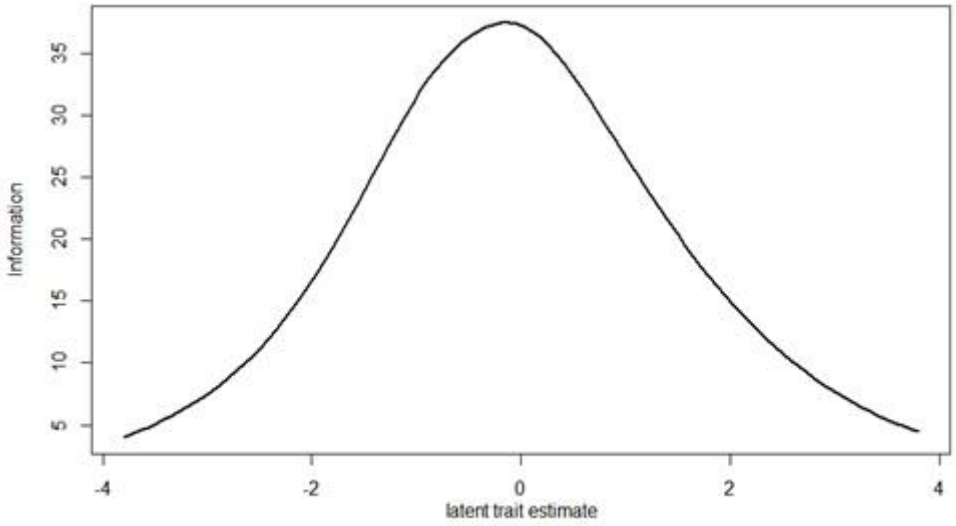

Figure 9. Test Information Function

4. Discussion

In the present study, the psychometric properties of the COPD-SIB were evaluated.

The investigation of the item bank was done by first executing an exploratory factor analysis and secondly a confirmatory IRT analysis. In the following, important findings will be summarized and put into relation with each other. Furthermore, limitations of the study, implications, and future perspectives will be elaborated.

4.1. Important findings

Overall, the results showed that the item bank is reasonably unidimensional. Inter-item correlations and factor analysis revealed that most of the items can be ascribed to one prevailing factor, which we label "discomfort due to COPD". However, some items performed poorly and showed weak or negative correlations. In order to determine how to treat these items, a comparison with the results of confirmatory analysis is necessary.

4.1.1. Comparison exploratory and confirmatory analysis

Items that were conspicuous in exploratory analysis also stand out in confirmatory analysis. Item 10 had a striking amount of negative correlations with other items. Likewise, this item did not load on the first factor. In confirmatory analysis, this item also performed badly. It had only poor discrimination and its b parameters had an illogical order. Likewise item 12, 13, 16 and 65 were conspicuous as they had poor correlations in factor analysis as well as poor discrimination in IRT analysis. Items 19, 53, 54, 56 and 66 were not assessable in IRT analysis and also performed badly in factor analysis. It can thus be concluded that these 10 items should not be included in the CAT. What is also conspicuous is that 7 out of these poorly performing items were items that are poled negative and hence had to be recoded prior to the analyses. Moreover, only half of the items had an appropriate discrimination parameter (33 out of 66 items).

4.1.2. Evaluation of the response categories

Another important finding is that 52 out of 66 items indicated that one of the categories was superfluous, or that categories were not logically ordered. This would suggest that fewer response options (as dichotomous) are more appropriate. In the SGRQ-C, most of the response categories were originally scored by means of a dichotomous true/false scale. In fact, this might also be more fitting for the COPD-SIB, as 17 out of the 25 items stemming from the SGRQ-C indicate that at least one category is superfluous.

Likewise, the 4 items stemming from the MRF-26 indicate superfluous response categories. These items are also originally scored by means of a dichotomous true/false rating scale.
Most of the ICCs indicate that 3 response categories would have been enough.

However the developers had clear reasons for choosing a polytomous response scale. Respondents in the cognitive interviews (cf. section 1.3.2.) frequently mentioned to prefer a polytomous response scale over a dichotomous true/false scale, since the possibility of giving only such a restricted answer - true or false - would restrict their desire to answer the items more flexibly. Therefore, the developers chose to use a polytomous response scale. Possible reasons for these findings will be discussed later on.

4.1.3. Item and Test information

A third important finding is the information rate covered by the items and the whole test.
As can be seen in figure 8, the present item bank covers a quite small range of theta values, ranging from $\Theta \approx -1$ to $\Theta \approx 0.5$. Figure 9 also shows that though the item banks measurement is most precise at $\Theta \approx 0$, it is quite peak. The two figures both emphasize that there is only weak measurement precision at $\Theta s < -1 > 0.5$.

4.2. Methodological defense

Edelen and Reeve (2007) argued that combining classical test analysis (as EFA) and IRT analysis (as with GPCM) serves as an adequate complementary method in the process of developing and evaluating an instrument. As they argue "insights from IRT analyses are most useful when they are complemented by a familiarity with the basic properties of the data from classical analysis"(p. 16). Hence, it was fairly reasonable to apply EFA in order to ensure that the COPD-SIB was sufficiently unidimensional. However, IRT-based item analysis has been shown to be advantageous over simply applying classical analysis, especially when developing CAT. As suggested by many authors (Weiss and Vale, 1987; Kubinger, 1993; Embretson and Reise, 2000), the two main advantages of IRT-based CATs is that they provide a) better test efficiency and economy and b) greater measurement accuracy. Especially for CAT, it is important to have information about each item which is as accurate as possible, since it aims to select those items, which provide the highest amount of information for each estimation of the measured latent trait.

4.2.1. Exclusion of items and cases

In exploratory analysis, several items and persons were excluded from further analysis. This can, of course, lead to reduction of the sample size, which, in turn, can harm measurement precision. Moreover, it was shown that exploratory analysis cannot deal with too many missing values in the dataset. Bernaards and Sijtsma (1999) argue, that if ”nothing is done about item non-response, this may highly influence results from factor analysis and other multivariate statistical analyses, since incomplete cases will simply be omitted from the data to prevent covariance matrices from not being positive (semi)definite“ (p. 278). Hence, two-way imputation was necessary.

A requirement for the executed two-way imputation was that there are not more than a 5.0% overall missing rate in the whole data set, as imputing more that 5.0% of the data would distort the picture. This can be concluded from the fact that the method of two-way imputation corrects for item as well as person effects. The imputed value is calculated using a mathematical formula including average scores person and item wise. These average scores are naturally computed by summing up all scores and dividing them by their total number. Too many missings would thus lead to a wrong estimation of the average score. Consequently, a wrong estimation of the value to be imputed will be derived. Hence, before imputing it is of crucial importance to avoid as much missing data as possible at earlier stages. However, there are no general rules available so far that state how many missing values a person or item might have before being excluded. In this thesis, it was thus chosen to follow the rules recommended by the developer of the syntax used for the imputation (van Ginkel, and van der Ark, 2010) who suggested removing respondents with more than 60.0% missing values.

4.2.2. The model used

The first consideration one has to make when choosing the most appropriate model is whether the data set has dichotomous or polytomous response categories. For polytomous items the Partial Credit Model (PCM), the Rating Scale Model (RSM), the Generalized Partial Credit Model (GPCM), the Graded Response Model (GRM) as well as the Nominal Model are available.

Secondly, one has to consider whether response categories are ordered or not. The later Nominal Model is only applicable for non-specific response order and, thus, it is not a suitable model for our analysis, as the response categories used in the COPD-SIB are ordered.

Next, choosing the right model is a question of how many parameters one wants to estimate. The number of parameters to estimate is also a question of sample size. Yu (2013) advises to use 1PL models in preference when there are sample size constraints. However, there is no clear evidence about sample size requirements for models with more parameters. While Tsutakawa and Johnson (1990) recommend a sample size of approximately 500 cases, later studies from Orlando and Marshall (2002) also suggest 200 or fewer cases can be an appropriate sample size for appropriate parameter estimation. Although the 1PL is a popular model to evaluate the psychometric properties of questionnaires (Haley, McHorney, and Ware, 1997; Raczek, Ware, Bjorner, Gandek, Haley, Aaronson NK, et al, 1998; Rost, Carstensen, and von Davier, 1997), in this thesis it was chosen to rather apply a 2PL model, since the examination of the discrimination ability for each item is crucial (regarding the development of a CAT). As can be seen in *Table 4*, the discrimination parameters substantially vary across the various items. Hence, we can assume that a 2PL fits the data better than a 1PL model would have, as the use of a 1PL model would have ignored this diversity in values of the discrimination parameter.

As RSM and PCM are 1PL models, only the GRM or GPCM remained. The choice between those two models is mostly due to personal preference, as they generally produce nearly identical results (Edelen and Reeve, 2007). This was also supported by research from Maydeu-Olivares, Drasgow, and Mead (1994), who applied ideal-observer technique to real-life polytomous models and found little difference in data fit between GRM and the GPCM. It should be noted that “a” and “b” parameters of GRM and GPCM cannot be compared directly as...

4.3. Limitations/future studies

As the purpose of the COPD-SIB is to be included in a 4 dimensional CAT, which will also consist of 3 other generic PROMIS item banks (cf. section 1.3.), it would be interesting to know how the items from the COPD-SIB behave in relation to the items from the other item banks. Due to limited time, this is beyond the scope of this thesis. Further studies should address this issue to ensure that the CAT model can be applied successfully.

Furthermore, the current design does not allow determining the specific causes of illogically ordered response categories. From the current point of view, only suggestions can be made. One possible explanation for these ambiguous findings might be the inclusion of a mid-point response option. As Alwin (2007) argues, inclusion of a mid-point response category leads to lower reliability. Additionally, finding from Hernández and colleagues

(2001) support this view. In their study, they carried out an IRT analysis. 8 out of 40 items on the pretest, and 19 out of 40 items on the posttest, which involved middle categories, did not show an ordered threshold. Another possible explanation may be what Jamieson (2004) states. According to him, terms such as "often" or "sometimes" may be ambiguous to the respondents and, thus, may result in inaccurate responses. However, a study from González-Romá andEspejo (2003) emphasizes the use of polytomous in favor of dichotomous response formats. Results of their study showed, that when compared to each other in terms of their information functions, the polytomous format performed better than the dichotomous one along the latent construct. Also several others suggest four to seven categories to be appropriate to obtain valid and reliable responses (Lozano, Garicai-Cueto, and Muniz, 2008; Weng, 2004). Future studies should address this problem more in detail as "rating scales are the communication medium between the researcher and survey" (Royal, Ellis, Ensslen, and Homan, 2010, p. 1) and should thus be as valid as possible.

A third limitation is that the chosen research strategy did not account for testing the choice of anchor items. As Peterson et al. (1982) states, the mean difficulty of anchor items should be close to that of total tests. It was beyond the scope of this thesis to address this requirement. However, the choice of which items should be chosen as an anchor items is crucial to the quality of equating (Sinharay and Holland, 2006). Hence, it is advisable for future studies to further examine the item bank according to this criterion. However, from the current point of view, it is reasonable to assume that the choice of anchor items did not influence the quality of equating too much, as they meet a second requirement states by von Davier, Holland and Thayer (2004). According to them, it is of importance for an anchor test to be representative that it should include the same content and statistical characteristics as the original version and thus be a miniature version of the original test. This is given by the current form of the COPD-SIB insofar, as the ten anchor items cover 6 of the given subdomains (cf. *Appendix A*). However, to ensure high quality of measurement, it is advisable to also include anchor items covering the two missing domains (symptoms, positive psychological functioning) in future studies.

4.4. Conclusion

Reviewing the item information curves for each item, several items should be rephrased or added. Poorly performing items could be rephrased in a way that the test will provide better information beyond the current peak ranging from $\Theta \approx -1 - 0.5$. On example for these items is item 9. This item was most informative at $\Theta \approx -1.5$. Likewise, item 63 could also serve as an

inspiring item, as it as it has a wide range of information for Θ from -0,5 to 1,5. Both of these items additionally provided good discrimination (item 9: a=0.857; item 63: a=0.963). However, they performed poor regarding their order of response categories. ICCs of both items displayed that one response category was superfluous. A example of an ideal item serves item 40, as it provides good information for Θ ≈ -1,5 and Θ ≈ 1, is highly discriminating (a= 1.719) and also its response categories are logically ordered.

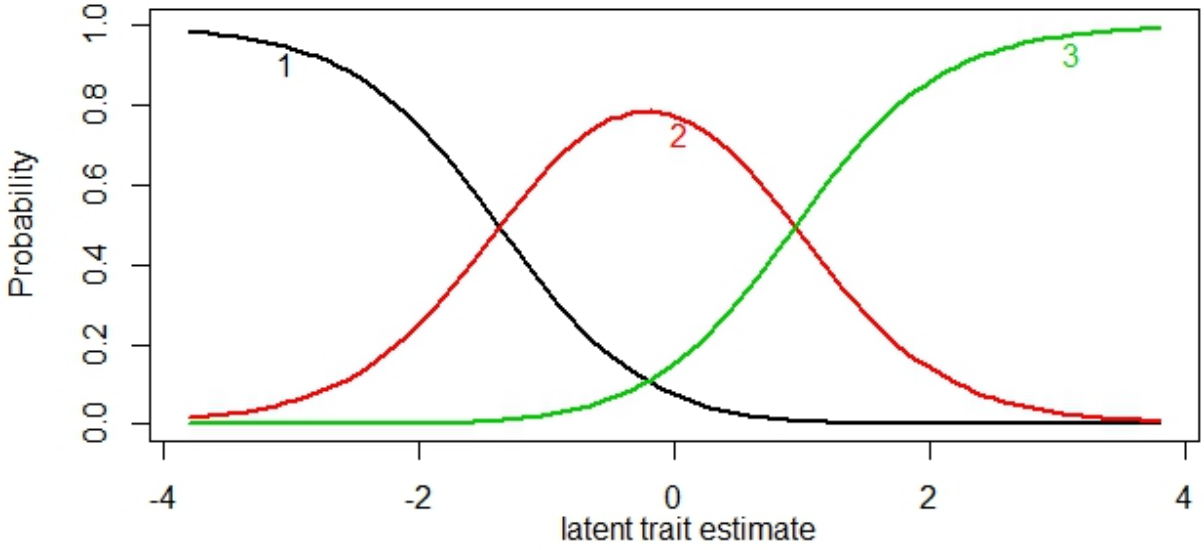

Figure 10. Item Response Category Characteristic Curve - Item 40

As item 9 and 63 indicate that out of four response categories one category is superfluous, while item 40 indicates that three categories are adequate, it can be concluded that items might be adjust in a way that response categories are reduced to only three categories (this will also be discussed more in detail in the following).

Regarding the item content of Item 9 (“because of my COPD, I feel unable to achieve my objectives”) item 63 (getting breathless when “getting washed or dressed”) and item 40 („I get tired easily”), these 3 items all have in common that they address the most *simplest* and *practical* problems of everyday life (do things, to wash oneself, to dress up, to become tired). It is strinking that items who performed poor in the analysis account for more *abstract* things in everyday life (e.g. item 24: “I value my life just as much as I did before I was diagnosed with COPD”, item 26: “I avoid thinking about how my COPD could get worse in the future”, item 16: „Since being diagnosed with COPD, I have lived more consciously”). It is thus reasonable that the item bank should be reviewed and abstract items should be rephrased in a more tangible sense.

Regarding the illogical and disarranged response categories, a possible solution might be to merge superfluous response categories. Linacre (2002), in this connection, suggests certain guidelines that should be followed when merging response categories. One of his guidelines states, that there should be ”at least 10 observations for each category“ as ”Each step calibration, Fk, is estimated from the log-ratio of the frequency of its adjacent categories“ (p.6) and thus low category frequency can lead to imprecisely estimated and potentially unstable results. This guideline was already applied in the present analysis, as IRT sufficient observations for each answering options in order to estimate the parameters accurately (cf. section 2.4.1.2.). Further guidelines can be found in *Optimizing Rating Scale Category Effectiveness* Linacre (2002). In their study, Royal, Ellis, Ensslen, and Homan (2010) applied these guidelines and could demonstrate that collapsing adjacent categories improves reliability. Also Smith, Wakely, de Kruif, and Swartz (2003) demonstrated that merging a 10-point response scale into a more meaningful 4-point scale provides a good way to improve measurement precision.

Regarding the findings that seven out of these poorly performing items were items that are poled negative and hence had to be recoded prior to the analyses, it can be inferred from the literature that these findings are not as surprising as one might think. To be clear: the item bank aims to measures *poor* HRQoL. Hence, “positive” poled items indicate *poor* HRQoL, while “negative” poled items indicate a *good* HRQoL. As most of the items were formulated to measure *poor* HRQoL the item bank appeared quite depressive to the respondents and they thus stated that also positive items should appear in the item bank. Hence, some items that indicate good HRQoL were included in the item bank, which had to be recoded for the analysis. As Bentler, Jackson and Messick (1971) state, sudden change in item wording may result is in remarkable difference in factor structure. One reason for this may be, as Schmitt and Stults name it, “careless response” (Schmitt and Stults, 1985). According to them, this does not mean that one responses randomly. In fact they hold, that “careless response” means that a respondent ,,is simply reading a few of the items in a measuring instrument, inferring what it is the items are asking of the respondent, and then responding in like manner to the remainder of the items in the instrument” (p. 367). Bentler et al. are describing two different types of acquiescence biases referring to these “careless responses”. First, there is the *agreement bias*. This response bias is referring to the tendency to agree (or answer positively) to an item regardless of the content of the questions. Respondents, thus just read over the negative poled question. Secondly, they mention the *acceptance acquiescence bias*. Here, a respondent endorse all items that they feel are true for oneself and concurrently disagree with

all items denying such characteristics. As the second type of bias is more relevant in personality measurements, only the first bias may account for the current COPD-SIB.

Anyhow, the researchers had reasonable motivations to include items that are poled negative, as respondents in the interviews (cf. section 1.3.2., step 1) often stated that they wish to also find items regarding positive psychological functioning in the item bank.

4.5. Suggestions

Concluding from this discussion, it would reasonable to suggest to adhere to a polytomous response scale (in favour of the desires stated by the respondents) but to reduce the number of response categories (in favour of a more precise outcome) and then to merge response categories retrospectively (also in favour of a more precise outcome). Moreover, though respondents stated to miss positive items, it is advisable to rephrase these items (again in favour of a more precise outcome).

4.6. Final Conclusion

In summary, it can be concluded that the item bank is yet insufficient to be included in the final version of the CAT. Before being included in the final CAT, the item bank should undergo several adjustments, as suggested above. Additionally, further research has to be initiated in order to examine this newer version of the adapted item bank.

5. References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* (Vol. 547). John Wiley & Sons.

Beattie, D. K., Golledge, J., Greenhalgh, R. M., & Davies, A. H. (1997). Quality of life assessment in vascular disease: towards a consensus. *European journal of vascular and endovascular surgery*, *13*(1), 9-13.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin,76*, 186-204.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research, 34*(3), p. 278.

Both, H., Essink-Bot, M. L., Busschbach, J., & Nijsten, T. (2007). Critical review of generic and dermatology-specific health-related quality of life instruments. *Journal of Investigative Dermatology*, *127*(12), 2726-2739.

Boucot, K. R. (1969). "Mrs. Young's 90th birthday party." *Arch Environ Health, 18*(3), 306.

Decramer, M., Janssens, W., & Miravitlles, M. (2012) Chronic obstructive pulmonary disease. *Lancet*, 3*79*, 1341–51.

DeMars, C. (2010). *Item response theory*. Oxford: University Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating*. New York: Springer.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5-18.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press

Fayers, P. M., & Machin, D. (2000). Item Response Theory and Differential Item Function. In: *Quality of Life: Assessment, Analysis and Interpretation*, 117-134. Wiley & Sons.

Ginkel van, J. R., & van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement, 29*(2), 152-153.

González-Romá, V., & Espejo, B. (2003). Testing the middle response categories"Not sure", "In between" and "?" in polytomous items. *Psicothema, 15*(2), 278-284.

Guyatt, G.H., Berman, L.B., Townsend, M., et al. (1987). A measure of quality of life for clinical trials in chronic lung disease. *Thorax 42*,773–778.

Hak, T., van der veer, K., & Jansen, H.(2004). The Three-Step Test-Interview (TSTI): An observational instrument for pretesting self-completion questionnaires, *ERIM Report ERS-2004-029-ORG*, Rotterdam: Erasmus Research Institute of Management.

Hernández, A., Espejo, B., González-Romá, V. & Gómez-Benito (2001). Likert-type response scales: is the response category «indifferent » relevant? *Metodología de Encuestas, 2*, 135-150.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549-559.

Hyland, M. E., Bott, J., Singh, S., & Kenyon, C. A. P. (1994). Domains, constructs and the development of the breathing problems questionnaire. *Quality of Life Research, 3*(4), 245-256.

Hyland, M. E., Singh, S. J., Sodergren, S. C., & Morgan, M. P. L. (1998). Development of a shortened version of the Breathing Problems Questionnaire suitable for use in a pulmonary rehabilitation clinic: a purpose-specific, disease-specific questionnaire. *Quality of life research, 7*(3), 227-233.

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., ... & DeWalt, D. A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research, 19*(4), 595-607.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education, 38*(12), 1217-1218.

Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., ... & Moinpour, C. M. (2015). Validation of the PROMIS physical function measures in a diverse US

population-based cohort of cancer patients. Quality of Life Research, 1-12. Doi: 10.1007/s11136-015-0992-9

Jones, P. W., Quirk, F. H., & Baveystock, C. M. (1991). The St George's respiratory questionnaire. *Respiratory medicine, 85*, 25-31.

Jones, P. W., Quirk, F. H., Baveystock, C. M., & Littlejohns, P. (1992). A self-complete measure of health status for chronic airflow limitation: the St. George's Respiratory Questionnaire. *American Review of Respiratory Disease*, *145*(6), 1321-1327.

Kubinger, K. (1993). Computerized Diagnostic-Psychometric Consideration. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*(3), 130-137.

Koller, M., Lorenz, W. (2002). Ziele des Heilens und Konzepte von Outcome in der modernen Medizin. *Recht und Politik im Gesundheitswesen, 8*(1),18–25.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurements, 3*(1), 85-106.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73-79.

Lopez, A. D., Shibuya, K., Rao, C., Mathers, C. D., Hansell, A. L., Held, L. S., ... & Buist, S. (2006). Chronic obstructive pulmonary disease: current burden and future projections. *European Respiratory Journal*, *27*(2), 397-412.

Maille, A. R., Koning, C. J. M., Zwinderman, A. H., Willems, L. N. A., Dijkman, J. H., & Kaptein, A. A. (1997). The development of the 'Quality-of-life for Respiratory Illness Questionnaire (QOL-RIQ)': a disease-specific quality-of-life questionnaire for patients with mild to moderate chronic non-specific lung disease. *Respiratory medicine*, *91*(5), 297-309.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among Paranletric Item Response Models for polychotomous ordered data. *Applied Psychological Measurement, 18*(3), 245-256.

McHorney, C.A., Haley, S.M., Ware, J.E. Jr. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology*, *50*(61).

Meguro, M., Barley, E.A., Spencer, S., Jones, P.W.(2007). Development and Validation of an Improved, COPD-Specific Version of the St. George Respiratory Questionnaire. *Chest, 132*(2), 456-63.

Mehta, T., Subramaniam, A. V., Chetter, I., & McCollum, P. (2003). Disease-specific quality of life assessment in intermittent claudication: review. *European journal of vascular and endovascular surgery*, *25*(3), 202-208.

Miller, M.D. (2002). Health-related quality of life. *Multiple Sclerosis Journal 8*(4), 269-270.

Mueller-Buehl, U., Engeser, P., Klimm, H. D., & Wiesemann, A. (2003). Lebensqualität als Bewertungskriterium in der Allgemeinmedizin. *ZFA- Zeitschrift für Allgemeinmedizin,79*(1), 24-27.

Muraki, E. (1997). A generalized partial credit model. In: *Handbook of modern item response theory*. New York: Springer.

Ninot, G., Soyez, F., & Préfaut, C. (2013). A short questionnaire for the assessment of quality of life in patients with chronic obstructive pulmonary disease: psychometric properties of VQ11. *Health and the Quality of Life Outcomes, 11*, 179. Doi:10.1186/1477-7525-11-179

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 14*(1), 50–59.

Paap, M. C., Bode, C., Lenferink, L. I., Groen, L. C., Terwee, C. B., Ahmed, S., Eilayyan, O., & van der Palen, J. (2014). Identifying key domains of health-related quality of life for patients with Chronic Obstructive Pulmonary Disease: the patient perspective. *Health and quality of life outcomes, 12*(1), 106. Doi: 10.1007/s11136-014-0860-z

Paap, M. C. S., Brouwer, D., Glas, C. A. W., Monninkhof, E. M., Forstreuter, B., Pieterse, M. E., & van der Palen, J. (2015). The St George's Respiratory Questionnaire revisited: a psychometric evaluation. *Quality of Life Research, 24*(1), 67-79.

Paap, M., Bode, C., Lenferink, L. I., Terwee, C. B., & Palen, J. (2015). Identifying key domains of health-related quality of life for patients with chronic obstructive pulmonary disease: interviews with healthcare professionals. *Quality of Life Research, 23*(10).

Pauwels, R. A., Buist, A. S., Calverley, P. M., Jenkins, C. R., & Hurd, S. S. (2014). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine, 163*(5).

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). *A test of the adequacy of linear score equating method*. In P. W. Holland & D. B. Rubin (Eds.),*Testequating,* New York: Academic Press, 71-135.

Raczek, A.E., Ware, J.E., Bjorner, J.B., Gandek, B., Haley, S.M., Aaronson, N.K., et al. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countri es: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*(14).

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Munster, Germany: Waxmann.

Royal, Ellis, Ensslen, & Homan (2010). Rating scale optimization in survey research: An application of the Rasch Rating Scale Model. *Journal of Applied Quantitative Methods, 5*(4), 586-596.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents?. *Applied Psychological Measurement, 9*(4), 367-373.

Sinharay, S., & Holland, P. (2006). Choice of anchor test in equating. *ETS Research Report Series, 2*, i-43. Princeton, NJ: Educational Testing Service.

Smith, E. V., Wakely, M. B., Kruif, R. E. de, Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63*, 369-391.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371-390.

Vidotto, G., Carone, M., Jones, P. W., Salini, S., & Bertolotti, G. O. (2007). Maugeri Respiratory Failure questionnaire reduced form: a method for improving the questionnaire using the Rasch model. *Disability & Rehabilitation, 29*(13), 991-998.

Walter, O. B., Becker, J., Fliege, H., Bjorner, J., Kosinski, M., Walter, M.,& Rose, M. (2005). Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT 1). *Diagnostica, 51*(2), 88-100.

Weiss, D. J., & Vale, C. D. (1987). Adaptive testing. *Applied Psychology, 36*(34), 249-262.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological measurement, 64*(6), 956-972.

World Health Organization (2008). COPD predicted to be third leading cause of death in. Retrieved from: http://www.who.int/respiratory/copd/World Health Statistics 2008/en/ accessed: February, 2015.

WHOQoL Group. (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological medicine, 28*(03), 551-558.

Wijkstra, P. J., TenVergert, E. M., Van Altena, R., Otten, V., Postma, D. S., Kraan, J., & Koeter, G. H. (1994). Reliability and validity of the chronic respiratory questionnaire (CRQ). *Thorax, 49*(5), 465-467.

Yu, C. H. (2013). A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling. Retrieved from: http://www.creative-wisdom.com/computer/sas/IRT.pdf. (accessed: May, 2015).

APPENDIX A.

COPD-specific Item Bank (Dutch)

Item	Thema	Anchor Item	Content
1	Situations triggering or enhancing breathing problems (S)		In gebouwen met air conditioning (bijvoorbeeld in het ziekenhuis)
2	(S)	√	Op koude dagen
3	(S)		Op mistige dagen
4	(S)		Op vochtige dagen
5	(S)		Op dagen met wind
6	(S)		In de buitenlucht in de pollentijd
7	(S)		Door boerderijdieren of huisdieren
8	(S)		Door bloemen, bomen, planten
9	Social	√	Door mijn COPD ben ik niet in staat alle dingen te doen die ik wil.
10	Coping	√	Ik heb er vertrouwen in dat ik met mijn COPD om kan gaan, ook al zouden de klachten verergeren.
11	Anxiety/hopelessness		Ik kan me voorstellen dat er mensen met zeer ernstige COPD klachten zijn die het leven niet meer de moeite waard vinden.
12	Autonomy	√	Ik vind het vervelend om hulp te moeten vragen van een ander, wanneer ik iets zelf niet kan.
13	Positive psychological functioning		Door mijn COPD waardeer ik mijn sociale contacten (bijvoorbeeld vrienden, partner en familie) meer.
14	Anxiety/hopelessness		Als ik nadenk over mijn COPD, krijg ik een uitzichtloos gevoel.
15	Coping	√	Ik vermijd activiteiten waarvan ik weet dat ik er vermoeid van raak.
16	Positive psychological functioning		Sinds ik COPD heb, sta ik bewuster in het leven.
17	Autonomy		Ik vind het frustrerend dat ik hulp aan moet nemen voor dingen die ik gewend was zelf te doen.
18	Anxiety/hopelessness		Als de klachten van mijn COPD verergeren, zie ik het leven niet meer zitten.
19	Coping		Ik ben tevreden met de dingen die ik nog kan.
20	Coping		Ik raak snel teleurgesteld wanneer iets me niet lukt door

			mijn COPD
21	Anxiety/hopelessness		Vanwege mijn COPD heb ik angst om alleen te zijn.
22	Coping		Als ik pieker over mijn COPD, vind ik het moeilijk daarover te praten.
23	Coping		Ik voel me beperkt door mijn COPD.
24	Positive psychological functioning		Mijn leven is nog even waardevol als voordat ik COPD kreeg
25	Coping		Ik vermijd activiteiten waarvan ik weet dat ik er benauwd van raak.
26	Coping		Ik vermijd nadenken over hoe mijn COPD in de toekomst zou kunnen verergeren.
27	Anxiety/hopelessness	√	Af en toe voel ik me zo benauwd/kortademig, dat ik bang ben dat ik zal stikken.
28	Coping		Ik vind het moeilijk om te accepteren dat ik door mijn COPD niet meer alles kan doen wat ik zou willen doen.
29	Emotions (official) - Autonomy (our label)		Ik heb last van het gevoel afhankelijk te zijn van anderen.
30	HRQoL impairment (official) - impact (label Nadine)		Door mijn COPD kan ik niet zo veel praten als ik zou willen.
31	Impact		Door mijn COPD heb ik soms geen controle over mijn ontlasting.
32	HRQoL impairment (official) - impact (label Lonneke)	√	Door mijn COPD ga ik minder dan gewoonlijk op bezoek bij vrienden of kennissen.
33	HRQoL impairment (official) - impact (label Lonneke)		Door mijn COPD breng ik veel meer tijd in mijn eentje door.
34	HRQoL impairment (official) - impact (label Lonneke)		Door mijn COPD heb ik liever dat iemand me vergezelt als ik buiten de deur ga.
35	Impact		Mijn hoesten is pijnlijk
36	Impact		Door mijn hoesten raak ik vermoeid
37	Impact		Ik raak kortademig wanneer ik praat
38	Impact		Ik raak kortademig wanneer ik mij voorover buig om iets op te pakken
39	Impact	√	Mijn hoesten of ademhalingsproblemen verstoren mijn slaap
40	Impact		Ik word snel moe

41	Impact	√	Ik schaam me als ik in het bijzijn van anderen ademhalingsproblemen heb of moet hoesten
42	Impact		Ik heb het gevoel dat mijn ademhalingsproblemen lastig zijn voor mijn omgeving (bijvoorbeeld vrienden, buren en familie)
43	Impact		Ik word bang of raak in paniek als ik niet genoeg adem kan krijgen
44	Impact		Ik heb het gevoel dat ik mijn ademhalingsproblemen niet onder controle heb
45	Impact		Ik ben zwak of minder valide geworden door mijn ademhalingsproblemen
46	Impact		Alles lijkt mij een te grote inspanning
47	Activity		Mijn ademhalingsproblemen maken het moeilijk om licht tuinwerk te verrichten (zoals wieden)
48	Activity		Mijn ademhalingsproblemen maken het moeilijk om te sporten (bijvoorbeeld hardlopen, tennissen of zwemmen)
49	Activity		Mijn ademhalingsproblemen maken het moeilijk om te dansen, golf te spelen of te bowlen
50	Coping		frustreerde het me dat ik niet meer alles kon doen wat ik wilde doen.
51	Anxiety/hopelessness		dacht ik soms wel eens: "het hoeft voor mij allemaal niet meer".
52	Coping		bleef ik het liefst de hele dag in bed/op de bank liggen als ik een ‘slechte’ dag had.
53	Coping		kon ik het accepteren, wanneer iets me niet meer lukte door mijn COPD.
54	Coping		ging ik ondanks dat ik door mijn COPD een activiteit niet meer goed kon uitvoeren, net zo lang door totdat het mij wel lukte.
55	Anxiety/hopelessness		raakte ik in paniek als ik moeilijk adem kon krijgen.
56	Coping		kon ik goed met mijn COPD omgaan
57	Coping		had ik mijn ademhalingsproblemen onder controle
58	Symptom		heb ik gehoest
59	Symptom		heb ik slijm opgegeven
60	Symptom		was ik kortademig

61	Symptom		heb ik last gehad van piepende ademhaling
62	Activity		Wassen of aankleden
63	Activity		Thuis rondlopen
64	Activity	√	Een wandeling maken
65	Activity		De trap opgaan (één verdieping)
66	Activity		Een steile helling oplopen
67	(S)		In gebouwen met air conditioning (bijvoorbeeld in het ziekenhuis)

APPENDIX B.

Overview Original Items, Scoring and Translation of COPD-Specific Item Bank

Item	Scoring	Original Items		Item changed to
		stem from	Content	
1	1 = Not at all 2 = A little bit 3 = Somewhat 4 = Quite a bit 5 = Very much	QoL-RIQ	Being in air-conditioned buildings	
2		QoL-RIQ	On cold days	
3		QoL-RIQ	On foggy days	
4		QoL-RIQ	On humid days	
5			Op dagen met wind	
6		QoL-RIQ	Being outside during the polling season	
7		QoL-RIQ	Due to domestic animals or pets	
8		QoL-RIQ	By flowers, trees, plants	
9	1 = Strongly disagree 2 = Disagree 3 = Neither agree nor disagree 4 = Agree 5 = Strongly agree	VQ11	I feel unable to achieve my objectives	Because of my COPD, I feel unable to achieve my objectives.
10			I am confident I will be able to cope with my COPD, even if the complaints get worse.	
11			I can imagine that there are people with severe COPD complaints, who feel that life is not worth living anymore.	
12			I don't like having to ask somebody to help me, when I cannot do something myself.	
13			Because of my COPD, I appreciate my social contacts (e.g., friends, partner, relatives) more.	
14			When I think about my COPD, I have a feeling of hopelessness.	
15			I shun activities I know will cause fatigue.	
16			Since being diagnosed with COPD, I have lived more consciously.	
17			I find it frustrating that I have to accept help for things I was used to doing myself.	

18			If my COPD symptoms get worse, I don't care about life anymore.	
19			I am content with the things I can still do.	
20			I feel disappointed, when I'm not able to do something because of my COPD.	
21			Because of my COPD I'm afraid of being alone.	
22			When I worry about my COPD, I find it hard to talk about it.	
23			I feel restricted, due to my COPD.	
24			I value my life just as much as I did before I was diagnosed with COPD.	
25			I shun activities I know will cause breathlessness.	
26			I avoid thinking about how my COPD could get worse in the future.	
27			Once in a while I have such severe shortness of breath that I fear I will suffocate.	
28			I find it hard to accept that I cannot do everything I would like to do, due to my COPD.	
29		QoL-RIQ	Feeling dependent upon others	I don't like the feeling of being dependent upon others.
30		MRF-26	Because of my lung disease, I cannot talk as much as I would like to.	Because of my COPD, I cannot talk as much as I would like to.
31			Because of my COPD, I am sometimes unable to control my bowel movements.	
32		MRF-26	Because of my COPD, I visit friends and acquaintances less frequently than I used to.	Because of my COPD, I go out to see friends or acquaintances less than usual.
33		MRF-26	Because of my COPD, I spend much more time alone.	Because of my COPD, I spend much more time alone.

34		MRF-26	Because of my COPD, I would like somebody to accompany me, when I go out	Because of my COPD, when I am outside I feel I need to have someone with me.
35		SGRQ-C	My cough hurts	
36		SGRQ-C	My cough makes me tired	
37		SGRQ-C	I am breathless when I talk	
38		SGRQ-C	I am breathless when I bend over	
39		SGRQ-C	My cough or breathing disturbs my sleep	
40		SGRQ-C	I get exhausted easily	I get tired easily
41		SGRQ-C	My cough or breathing is embarrassing in public	
42		SGRQ-C	My chest trouble is a nuisance to my family, friends or neighbours	I feel that my chest trouble is a nuisance to my environment (e.g. family, friends or neighbours)
43		SGRQ-C	I get afraid or panic when I cannot get my breath	
44		SGRQ-C	I feel that I am not in control of my chest problem	
45		SGRQ-C	I have become frail or an invalid because of my chest	
46		SGRQ-C	Everything seems too much of an effort	
47		SGRQ-C	My breathing makes it difficult to do things such as walk up hills, carrying things up stairs, light gardening such as weeding, dance, play bowls or play golf	My breathing problems make it difficult to do light gardening, such as weeding.
48		SGRQ-C	My breathing makes it difficult to do things such as carry heavy loads, dig the garden or shovel snow, jog or walk at 5 miles per hour, play tennis or swim	My breathing problems make it difficult to exercise (e.g., jogging, playing tennis, or swimming).
49		SGRQ-C	My breathing makes it difficult to do things such as walk up hills, carrying things up stairs, light gardening such as weeding, dance, play bowls or play golf	My breathing problems make it difficult to do things such as dancing, playing golf, or playing bowls.
50			It frustrated me that I couldn't do everything I wanted to do anymore.	
51			I thought sometimes, I'm	

	1 = Never 2 = Rarely 3 = Sometimes 4 = Often 5 = Always		really fed up with everything.	
52			I wanted to stay in bed/ lie down on the couch all day, when I had a "bad" day.	
53			I could accept it, when I was not able to do something anymore, due to my COPD.	
54			I persevered until I had finished an activity, despite the fact that I couldn't perform that activity well, due to my COPD.	
55			I panicked, when I had trouble breathing	
56			I could cope with my COPD.	
57			I got my breathing problems under control.	
58		SGRQ-C	I cough: mosty days a week/several days	I coughed.
59		SGRQ-C	I bring up phlegm (sputum): mosty days a week/several days	I brought up phlegm (sputum).
60		SGRQ-C	I have shortness of breath: mosty days a week/several days	I had shortness of breath.
61		SGRQ-C	I have attacks of wheezing: mosty days a week/several days	I had attacks of wheezing.
62		SGRQ-C	Getting washed or dressed	
63		SGRQ-C	Walking around the home	
64		SGRQ-C	Walking outside on the level	Going for a walk
65		SGRQ-C	Walking up a flight of stairs	Walking up a flight of stairs (one floor)
66		SGRQ-C	Walking up hills	

APPENDIX C.

Syntax used in R-statistics

```
setwd("D:\\VoorIRTdata")

library(foreign)

mydata <- read.spss("boekjesmerged.sav",use.value.labels = FALSE, to.data.frame=TRUE)

myMatrix <- data.matrix(mydata)

myMatrix2 <- myMatrix[,2:67]

library(ltm)

selectionofitems <- myMatrix2[,c(1:66)]

outputNadine.2PL <- gpcm(selectionofitems, constraint = c("gpcm"), IRT.param = TRUE,
start.val = NULL, na.action = NULL, control=list(iter.qN=2000, GHk=19))

outputNadine.2PL

plot(outputNadine.2PL, type = "ICC", items = 1, lwd = 2, xlab="latent trait estimate")

plot(outputNadine.2PL, type = "IIC", items = 3, lwd = 2, xlab="latent trait estimate")

plot(outputNadine.2PL, type = "IIC", items = 1:66, lwd = 2, xlab="latent trait estimate")

plot(outputNadine.2PL, type = "IIC", items = 0, lwd = 2, xlab="latent trait estimate")
```